Multimed Tools Appl (2009) 42:57–71
DOI 10.1007/s11042-008-0251-y

# Can feature information interaction help for information fusion in multimedia problems?

**Jana Kludas · Eric Bruno · Stéphane Marchand-Maillet**

Published online: 15 November 2008
© Springer Science + Business Media, LLC 2008

**Abstract** This article presents the information-theoretic based feature information interaction, a measure that can describe complex feature dependencies in multivariate settings. According to the theoretical development, feature interactions are more accurate than current, bivariate dependence measures due to their stable and unambiguous definition. In experiments with artificial and real data we compare first the empirical dependency estimates of correlation, mutual information and 3-way feature interaction. Then, we present feature selection and classification experiments that show superior performance of interactions over bivariate dependence measures for the artificial data, for real world data this goal is not achieved yet.

**Keywords** Multi modal information fusion · Feature selection and construction · Feature information interaction

## 1 Introduction

With the rise of Web 2.0 and its tendency to be populated with an ever-increasing amount of images and videos, multimedia processing has become a lively discussed field in research. At the center of multimedia systems, there is an essential need for information fusion due to the multi modal nature of its data. Hence, the fusion of multi modal data (e.g. text and images) has a large impact on the performance of

J. Kludas (✉) · E. Bruno · S. Marchand-Maillet
CUI, University of Geneva, 7 route de Drize, 1227 Carouge, Switzerland
e-mail: jana.kludas@unige.ch

E. Bruno
e-mail: eric.bruno@unige.ch

S. Marchand-Maillet
e-mail: marchand@cui.unige.ch

algorithms for multimedia indexing, retrieval and classification, object recognition as well as for data preprocessing like feature selection or data model development.

Information fusion has established itself as an independent research area over the last decades, but a theoretic framework to describe general information fusion systems is still missing [12]. Still today, the understanding of how fusion works and by what it is influenced is limited. In multimedia document retrieval, especially for web applications, the visual component is still lacking behind expectations. This can be seen for example in the INEX 2006 [24] and 2007 Multimedia Tracks, where text-based runs outperformed all others. Other examples of text-dominant retrieval strategies are the commercial image search engines by Google, Yahoo!, that do not use any visual features.

The work done so far on information fusion in multimedia settings can be divided into two main tracks: (1) fusion of independent or complementary information by assuming or creating independence and (2) fusion of dependent information by exploiting their statistical dependencies. Both are applied in multimedia processing problems equally successfully. Neither of these approaches is superior.

Aligned to the second approach, we investigate here another way of analyzing input data for multimedia problems based on feature information interactions with the long term goal of improving the performance of multimedia document retrieval and classification. This multivariate, information theoretic based dependence measure is more accurate in detecting the hidden data structures e.g. situations, where the independence assumption is sufficient and where the dependency between the input data is not negligible. Freitas [7] gives a broad overview about the importance of feature interaction in data mining. We think that the processing of and especially the fusion in large scale multimedia collections is affected by similar problems and that interactions can help to overcome them.

In Section 2, we discuss in more detail state-of-the-art fusion approaches with independent and dependent input data and their shortcomings. Next, we present in Section 3 the idea of feature interaction information and how it can help to improve information fusion algorithms. In Section 4, we give the results of data analysis experiments with artificial and real data, as well as results of classification experiments based on feature selection and classification. Finally, conclusions and future work are given in Section 5.

## 2 Related work

Our article discusses the problem of information fusion, but most of the related work can be found in multimedia processing where information fusion is only implicitly treated as one part of the problem. We review some example approaches and explain when and why they may fail.

In early years of information fusion research, scientists fused different information sources by assuming independence between them. One of the first works on classifier and decision fusion used this principle, where they fusied neural network outputs [20]. The independence assumption is still widely used in machine learning as for example in the nave Bayes classifier. Its success is based on its simplicity in calculation and the learned models, as well as its robustness in estimating the evidence [10]. Approaches that fuse independent or complementary sources mostly belong

to classifier and decision fusion, where each modality of the input is first processed separately and then a final decision is based on the individual results. This principle has been applied e.g for multimedia retrieval [13, 26], multi modal object recognition [25], multi-biometrics [19] and video retrieval [27].

Despite the successful application of this approach for some problems, it seems to fail completely for others. In [19], it is shown that the violation of the independence assumption hurts the information fusion performance. So a trade-off between simple and fast calculated results and their accuracy is necessary. That loss in performance was empirically explained in [5], where the authors showed that the maximum performance in their multi-biometrics application can be achieved only if the statistical dependencies between the modalities are taken into account. These algorithms are also called myopic, because they treat all attributes as conditionally independent given the class label [14].

To circumvent the problem of attribute dependencies in data, other approaches try to create independence with the help of linear transformation methods like principal and independent component analysis (PCA/ICA), factor analysis and projection pursuit as reviewed in [8]. Unfortunately, these methods are not sufficient to eliminate all dependencies in the data, since they target only pairwise and linear feature dependencies [21]. In addition, the authors showed empirically that their multi modal object recognition problem is affected by higher order dependency patterns. A similar result was found in [22]. In the multimedia classification task the Support Vector Machine (SVM) approach using an ICA-based feature selection was outperformed by a SVM on the original data set.

Multimedia processing approaches that explicitly exploit attribute dependencies fuse the information preferably at data or feature level. Example applications are multimedia summarization [2], text and image categorization [4], multi modal image retrieval [23] and web document retrieval [28]. These approaches all exploit some form of attribute dependency at data level like co-occurrence (LSI [15]), correlation (kCCA [22]) or mutual information. As examples for late fusion, where classifier dependencies are exploited, can be named copula functions [9] or nonlinear fusion algorithms based on SVM [3].

The most important shortcoming of those algorithms is that they only take bivariate dependencies into account, even though they work in a multivariate setting [16]. High level feature relationships such as conditional dependencies of a feature pair to a third variable e.g. the class label are neglected. For now there exists no proof that these higher order dependencies have an impact on the performance of multimedia processing systems, but in [17] the exploitation led to a performance improvement.

## 3 Feature information interaction

Before the introduction of feature interaction by Jakulin and Bratko [10] there was no unifying definition of feature dependence in multivariate settings, but similar formulae have emerged independently in other fields from physics to psychology. Feature information interaction or co-information as it was named in [1] is based on McGill's multivariate generalization of Shannon's mutual information. It describes the information that is shared by all of $k$ random variables, without over counting

redundant information in attribute subsets. So, it finds irreducible and unexpected patterns in data that are necessary to learn from data [18].

This general view of attribute interactions could help machine learning algorithms to improve their performance. For example attribute interactions can be helpful in domains where the lack of expert knowledge hinders the selection of very informative attributes sets by finding interacting attributes needed for learning. Another example is the case when the attribute representation is primitive and attribute relationships are more important than the attributes themselves. Then similarity based learning algorithms fail, because the proximity in the instance space is not related to classification in this domain.

Two levels of interactions can be differentiated: (1) relevant non-linearities between the input attributes, which are useful in unsupervised learning and (2) interactions between the input attributes and the indicators or class labels, which is needed in supervised learning.

The k-way interaction information as found in Jakulin and Bratko (unpublished manuscript) for a subset $\mathcal{S}_i \subseteq \mathcal{X}$ of all attributes $\mathcal{X} = \{X_1, X_2, ..., X_n\}$ is defined as:

$$I(\mathcal{S}) = -\sum_{\mathcal{T}\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{T}|} H(\mathcal{T}) = I(\mathcal{S}\setminus X|X) - I(\mathcal{S}\setminus X), X \in \mathcal{S} \tag{1}$$

with the entropy defined as:

$$H(\mathcal{X}) = -\sum_{\mathcal{X}\in\mathcal{S}} P(X) log_2 P(X), \tag{2}$$

where in case of several variables the joint probability distribution is used. The feature interaction for $k = 1$ reduces to the single entropy, for $k = 2$ to the well known mutual information and for $k = 3$ attributes $A, B, C$ to McGill's multiple mutual information:

$$I(A; B) = H(A) + H(B) - H(A, B) \tag{3}$$

$$\begin{aligned} I(A; B; C) &= I(A; B|C) - I(A; B) \\ &= H(A, B) + H(A, C) + H(B, C) \\ &\quad -H(A) - H(B) - H(C) - H(A, B, C). \end{aligned} \tag{4}$$

According to this definition 3-way information interaction will be only zero iff $A$ and $B$ are conditionally independent in the context of $C$, because then $I(A; B|C) = I(A; B)$. So it gives only the information exclusively shared by the involved attributes. Information interactions are stable and unambiguous, since adding new attributes changes not already existing interactions, but adds only new ones. Furthermore they are symmetric and undirected between attribute subsets.

An important characteristic of $k$-way feature information interactions with $k > 2$ is that it can result in positive and negative values. Normally, when we consider Markov chains $A \to B \to C$, the data processing inequality states that conditioning always reduces the information $I(A; B|C) \leq I(A; B)$. This way the

3-way mutual information would be limited to $I(A;B;C) \leq 0$. But the problem of feature interactions is not a Markov chain, that is why it is possible that $I(A;B;C) > 0$. For example, let $C = A + B$ and let $A$ and $B$ be independent random variables, then $I(A;B) = 0$, but $I(A;B|C) = H(A|C) - H(A|B,C) = P(C = 1)H(A|C = 1) = 0.5bit$. The variables $A, B$ are said to have a synergy towards $C$. Thus, we can distinguish two different types of feature information interactions:

### 3.1 Synergy $I(A;B;C) > 0$

In case of positive interactions the process benefits from an unexpected synergy in the data. In statistics this phenomena is called moderating effect and has been known for a long time. Synergy occurs when $A$ and $B$ are statistical independent, but get dependent in the context of $C$ as can be seen in Fig. 1a. Myopic feature selections are unable to exploit the synergy in the data.

### 3.2 Redundancy $I(A;B;C) < 0$

Negative interactions occur when attributes partly contribute redundant information in the context of another attribute, which leads to a reduction of the overall dependence. It is shown in Fig. 1b on behalf of the redundant attributes $A, B$ towards a third attribute $C$. In supervised learning the negative influence of redundancy can be resolved by eliminating unneeded redundant attributes, but it could be advantageous in unsupervised learning in the case of noisy data.

In any case myopic voting function that are based on the independence assumption as well as fusion algorithms that use only local dependencies are confused by positive and negative feature interactions, which results in decreased information fusion performance.

In the following section we compare empirical estimates of correlation, mutual information and 3-way feature information interaction for artificial and real multi modal data to draw conclusions about their usefulness as dependence measure in information fusion.

### 3.3 Calculation of 3-way feature information interactions

The calculation of the full feature information interaction matrices is expensive, since the size of all possible combinations is $M^k$, where $k$ is the size of the feature subset

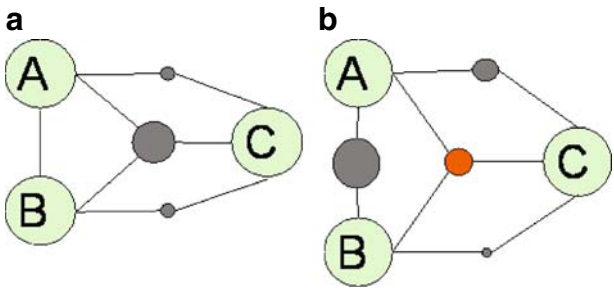



**Fig. 1** Interaction diagrams of different types of information interactions between $A$, $B$ and $C$ (**a**, **b**)

and $M$ the number of features. We formalize our problem as follows: $d = [1, ..., N]$ are the documents, $f_d = [1, .., M]$ are the extracted features and $l_d = [1, ..., C]$ are the class labels, that are given as ground truth. The features and labels are represented as probabilities over the documents such as:

$$P\left(f^i_{d_j}\right) = \frac{m_{f_i}}{m_{d_j}}$$

$$\sum_i P\left(f^i_{d_j}\right) = \quad 1 \quad \forall d_j. \tag{5}$$

and

$$P\left(l^i_{d_j}\right) = \begin{cases} 1, & d_j \in c_i \\ 0, & otherwise \end{cases} \tag{6}$$

where $m_{f_i}$ is the number of occurrences of a feature in a document and $m_{d_j}$ the number of all features occurring in a document. These descriptions are conform to the frequentist interpretation of probability and result in the discrete probability matrix $P(F)$ of size $[M \times N]$ and $P(L)$ of size $[C \times N]$.

In the following experiments, we applied a sub-sampling strategy to approximate interactions with $k = 3$ between two features and the class label. Using a normal random distribution, we chose to draw a very small sample set $M_S << [M \times M \times C]$ from the original search space. To do so, we approximated the joint entropies of the features and class labels by contingency tables. Then, the approximated feature interaction $I_S(A; B; C)$ between two random features and the class label is calculated as described in the previous subsection.

This is in any case a sub optimal solution, since a lot of significant interaction will be missing. Furthermore, also a lot of redundant interaction values are calculated, which is due to the symmetry of the interactions.

## 4 Experiments

For the objective evaluation of the different dependence measures, we first conducted tests on simple artificial data sets, where the relations between the input variables as well as their relations towards the class labels are known. Then, we applied the analysis to a real world collection: DB of the University of Washington (UW), Seattle (http://www.cs.washington.edu/research/imagedatabase/groundtruth/). Finally, we conducted classification experiments on the Washington collection using feature selection and construction based on the different dependence measures.

### 4.1 Feature relationship analysis on artificial data

The first artificial data set is based on an AND combination of three binary, random variables that define one of the three classes such as $l^1 = 1$ if $\left(f^1 \bigwedge f^2 \bigwedge f^3\right)$. Class 2 and 3 are respectively dependent on features $f^4$, $f^5$, $f^6$ and $f^7$, $f^8$, $f^9$. So, the intra-class relation between the variables is dominated by redundancy. The variables that depend on different classes, also referred to as the inter-class relation, are independent from each other. We generated $N = 10000$ samples as documents.

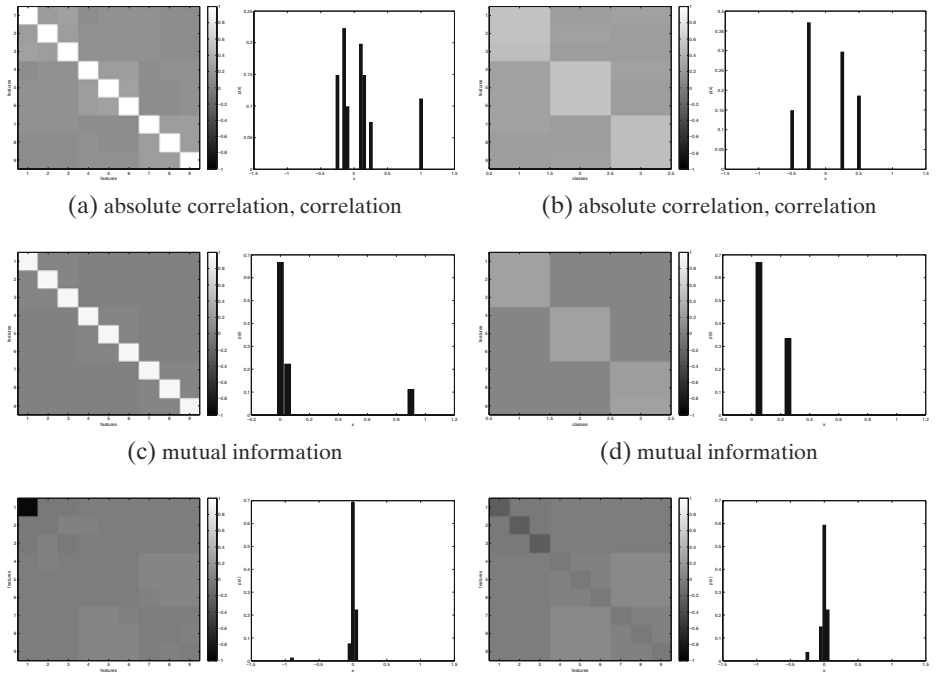
The statistical correlation is calculated by means of the Pearson correlation coefficient, which gives the amount of the linear coherence between two random signals $A$ and $B$:

$$Cor(A, B) = \frac{cov(A, B)}{\sigma_A \sigma_B}. \tag{7}$$

For the unsupervised case, which describes the correlation between the input features, a correlation matrix $Cor(F, F^T)$ of size $[M \times M]$ is built. The supervised case represents the correlation between the features and the class labels $Cor(F, L^T)$ and results in a correlation matrix of size $[M \times C]$. Mutual information and 3-way feature interactions (here a calculation of the full interaction matrix was possible) are calculated as described in the last section.

Figure 2 shows the empirical estimates and histograms of the correlation matrix, the mutual information and the 3-way information interaction respectively for the unsupervised (features towards features) and the supervised (features towards class labels) case. In both cases, all dependence measures succeed in finding the 3 dependent intra-class variables, but with differences in accuracy.

Correlation, for example, is constantly overestimating the dependencies, because it shows no independence for the inter-class variables. Furthermore, the knowledge of positive or negative correlation are not useful for information fusion, but only the absolute magnitudes.



(a) absolute correlation, correlation (b) absolute correlation, correlation

(c) mutual information (d) mutual information

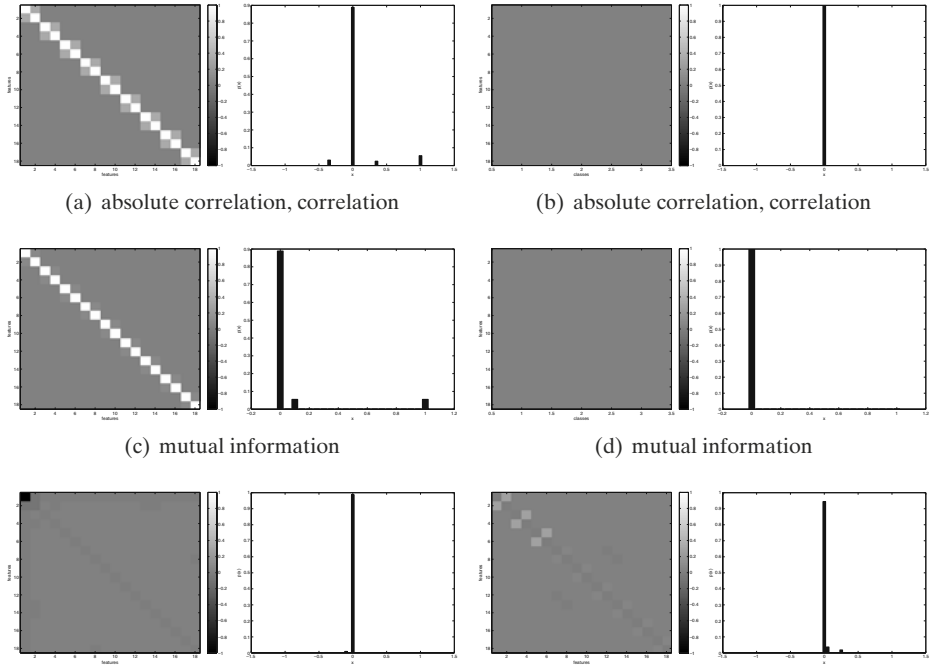
(e) 3-way interaction ($f^1 = 1$) (f) 3-way interaction ($I^1 = 1$)

**Fig. 2** Unsupervised (**a**, **c**, **e**) / supervised (**b**, **d**, **f**) case for AND combined artificial data

Mutual information performs similarly in accuracy as information interaction. So, it finds the inter-class independence of the input variables as well as the dependence of the intra-class variables. Finally, information interaction is giving the most detailed information about the data structure. For the intra-class variables, it results in negative interaction, which indicates redundancy. The inter-class information interactions are zero. Surprisingly positive interactions, hence synergy, appears between the blocks of intra-class variables, where we are not sure yet how to explain this.

The second and more interesting artificial data set is based on the AND set, but now each input variable is replaced by its XOR combination of two variables. It has again three classes, where each depends now on six input variables. This new data set is therefore a parity problem, which contains synergy between the XOR combined variables and their class labels, whereas the AND set only contains redundancy. We generated again $N = 10000$ samples as documents $d$, that are described by $M = 18$ features and define $C = 3$ classes.

Figure 3a, 3c and 3e show the empirical estimates and the histograms for the unsupervised case. Correlation finds independence between all variables except between the parity variables, where it results randomly in positive or negative correlations. This can be seen in the histogram that gives not the absolute values. Mutual information as well as the 3-way information interaction also show only the dependence between the parity variables. So none of the investigated dependence measures finds the features that one class depends on in the unsupervised setting.



(a) absolute correlation, correlation (b) absolute correlation, correlation

(c) mutual information (d) mutual information

(e) 3-way interaction ($f_3 = 1$) (f) 3-way interaction ($c = 1$)

**Fig. 3** Unsupervised (**a**, **c**, **e**) / supervised (**b**, **d**, **f**) case for OR combined artificial data

The results of the supervised case, that are presented in the Fig. 3b, 3d and 3f, show a clear advantage of information interaction over the other dependence measures. Correlation and mutual information do not succeed to find the parity variables because they are based only on bivariate relationships. Information interaction finds correctly synergy between the parity variables and detects all dependent variables of a class (in Fig. 3f the example for class 1 is given).

To summarize, it can be said that feature information interactions can detect complex dependence structures in data sets by giving their irreducible patterns. This is especially true for parity problems. Furthermore, it allows to differentiate feature relationships into synergies and redundancies, which we feel is useful knowledge to exploit in information fusion for multimedia systems.

### 4.2 Feature relationship analysis on real data

For the real data experiments, we used the Washington collection, which consists of $N = 886$ documents, which are images annotated with 1 to 10 keywords. They are grouped into $C = 20$ classes like for example football, Barcelona and Swiss mountains. The extracted feature set $F$ consists of the global color and texture histograms which result in $M_c = 165$ and $M_x = 164$ features respectively. Additionally, we constructed a textual feature vector of size $M_t = 297$ with the term frequencies of the keywords. The continuous variables are discretized with a simple equal length quantizer.

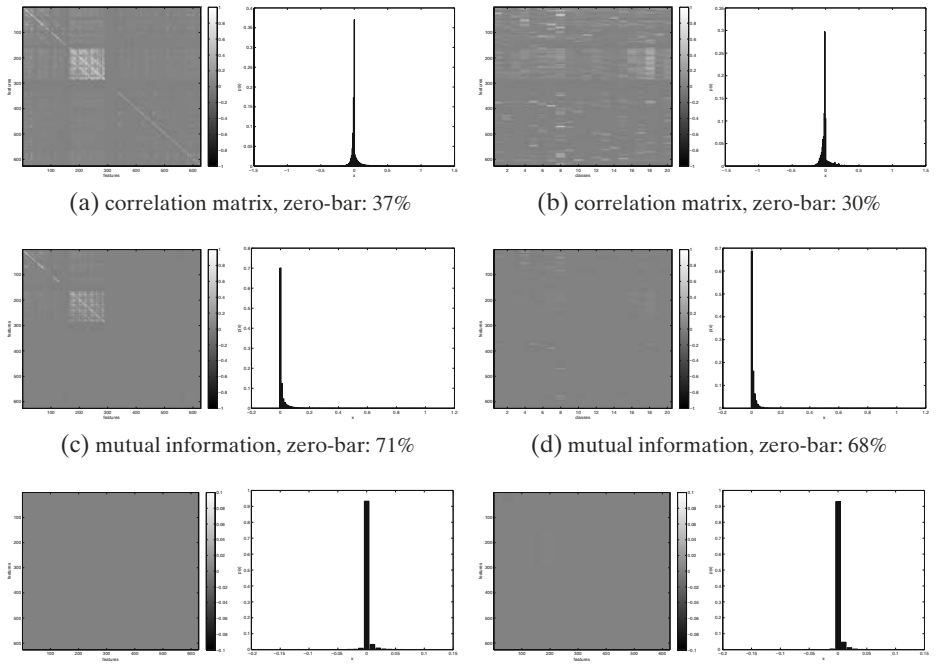
Ignoring the class labels, we first investigated the feature dependencies for the unsupervised setting. As described above, the full calculation of the 3-way feature interactions is infeasible. So we largely under-sample the search space of size $I(F_{k=3}) = [626, 626, 626]$ by calculating only interactions for a randomly selected subset of size $M_S = 80000$. The same is done for the supervised case $I(F_{k=2}, L) = [626, 626, 20]$. One should keep in mind that all the following results are based on incomplete interaction matrices.

Figure 4a, 4c and 4e give the empirical estimates of the dependence measures and their histograms. As expected the feature information interactions show only little dependence in the feature set. Be aware that the interaction diagrams are scaled between $[-0.1, 0.1]$ compared to $[-1, 1]$ for correlation and mutual information. So, it is clearly visible that the latter two, both 2-way dependence measures, indicate much higher relationships (in number and magnitude) between the features. Hence, one can state that they also overestimate the feature's dependencies for real data sets as they do for the artificial data sets.

The results for the supervised setting are shown in Fig. 4b, 4d and 4e. Again, the scale of the information interaction diagrams is set to $[-0.1, 0.1]$. Here the correlation between the features and their class labels results in high dependencies that are neither supported by the mutual information nor the 3-way feature information interaction. Mutual information overestimates slightly the dependencies.

### 4.3 Classification experiments with feature selection and construction

In this section, we present classification experiments that compare our approach of feature selection and construction based on feature information interaction to a baseline system that uses no feature selection and systems that do a feature selection



(a) correlation matrix, zero-bar: 37%

(b) correlation matrix, zero-bar: 30%

(c) mutual information, zero-bar: 71%

(d) mutual information, zero-bar: 68%

(e) 3-way interaction ($fv_3 = 1$) zero: 95%

(f) 3-way interaction ($cl = 1$) zero: 93%

**Fig. 4** Unsupervised (**a**, **c**, **e**) / supervised (**b**, **d**, **f**) case for the Washington collection

based on correlation, mutual information and random selection. The goal is to determine whether the knowledge of synergies and redundancies can improve a classification task in multimedia setting and how it is best exploited. The test collection is again the Washington collection.

The classification is done with the SVM light library [11] using a RBF kernel. We followed two different strategies: (1) a feature selection with late or hierarchical fusion over each modality and (2) a feature construction implemented by a late fusion over feature subsets of size $k = 3$. We ran a cross validation to optimize the parameters of the SVM. As training set we randomly selected for each run and class 5 positive and 7 negative examples, the rest of the examples were used as test set. The experiments were run with the one-against-all classification strategy, where the classification errors were averaged over all classes and over 10 runs.

**Feature selection** A simple, but nevertheless common strategy for feature selection approaches, that exploit statistical dependencies in a supervised setting, is the calculation of the pair wise relationships between the attributes and the class labels [6]. From the features that were ranked in descending order according to the applied dependence measure the best $M_s$ are selected as the subset $f_s = [1, .., M_s]$ to be included in the classification.

Our dependence measure of 3-way feature information interactions $I(F_{k=2}, L)$ divides into the absolute value $argmax|I(F_{k=2}, L)|$ labeled as (abs), the synergies $argmax(I(F_{k=2}, L))$ (syn) and the redundancies $argmin(I(F_{k=2}, L))$ (red). As

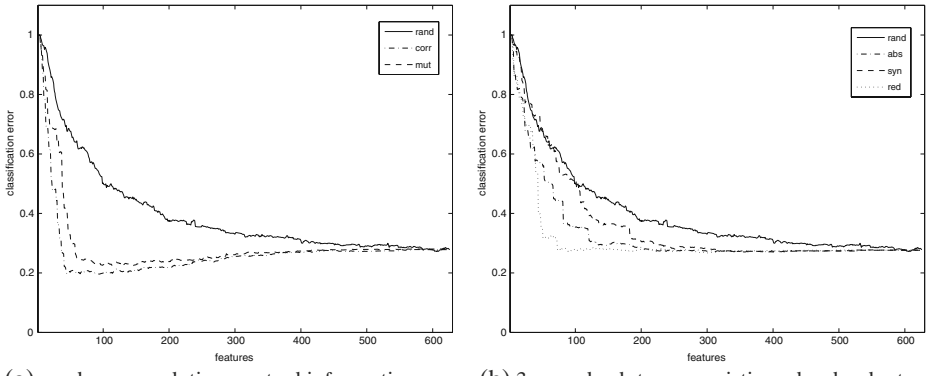
dependency measures to compare our approach with, we chose absolute correlation $argmax|Cor(F, L^T)|$ (corr), mutual information $argmaxI(F, L)$ (mut2D) and random selection (rand). Figure 5 shows the developing of the classification error over the number of selected features $M_s$ for correlation, mutual information and random selection on the left hand side and redundancy, synergy and absolute interaction on the right hand side.

As can be easily seen in the plot, the correlation based feature selection performs best with $e = 0.19$ at $M_s = 46$. Until about $M_s \sim 300$ it outperforms the baseline based on all features significantly. Thereafter, the systems with the feature selection based on mutual information and redundancy follow with a similar performance with $e = 0.22$ with $M_s = 100$ and $e = 0.27$ at $M_s = 77$ respectively. They can only slightly outperform the all feature baseline, that achieves a classification error of $e = 0.28$. The systems based on absolute information interactions, synergistic features and random selection are completely unfeasible.

We conclude from this experiment that, first of all, the baseline can be outperformed by feature selection based on correlation between the features and the class label. Concerning the performance of our approaches based on information interaction, we conclude that with feature selection only the redundancy information can be exploited, but this is not sufficient to attain an equal performance as the baseline or the correlation based feature selection.

**Feature construction** It is for this reason that we tried a simple feature construction approach on the same data and feature evaluation measures. It is set up again as a hierarchical SVM. But now, we create in the first step a mid-level feature over each synergistic, redundant or correlated feature subset by using a SVM. The results are then fused in a second level SVM towards the final classification result. In this way we select highly synergistic, redundant or correlated feature subsets towards the class labels.

The feature construction results based on correlation (corr), redundancy (red) and synergy (syn) are shown in Fig. 6 for one to hundred feature subsets. Now the synergistic features outperform largely the redundant ones with $e = 0.32$ at only




(a) random, correlation, mutual information

(b) 3-way absolute, synergistic and redundant interaction

**Fig. 5** Classification errors for the Washington collection with feature selection (**a**, **b**)

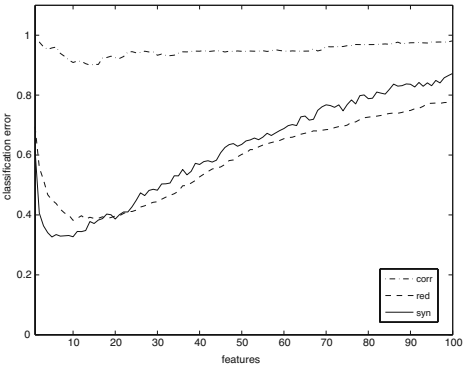



**Fig. 6** Classification errors for the Washington collection for fusion at feature subset level

(a) 3-way synergistic and redundant interaction, correlation

$M' = 5$ feature subsets of size $k = 3$, hence it uses only 10 features plus the class label. Still, the synergistic feature subsets stay behind the performance of the full feature set, but it achieves an acceptable classification result with only 1/100 of the original feature set. This is the steepest reduction of the classification error within the first few features, which makes this strategy valuable for extreme feature selection.

The experiments show, first of all, that feature redundancy and synergy have to be treated separately in information fusion. The first one is best exploited with feature selection and the other one achieves better results with feature construction. Pair wise dependence measure can not detect this difference in feature dependencies. We think that these preliminary results are promising and that further research towards an efficient and complete calculation of feature information interaction, that also allows the computation of higher order interactions $k > 3$, and the specialized treatment of synergies and redundancies can help to improve information fusion in future multimedia systems.

## 5 Conclusions and future work

The article reviews the formal theory and characteristics of feature information interaction, an information-theoretic dependence measure. Through its stable and unambiguous definition of feature relationships it can more accurately determine dependencies, because e.g. redundant contributions to the overall relationships are not over counted. Experiments on artificial data, where the feature dependencies are known, undermine the theoretically claimed superior performance of information interactions over bivariate dependence measures like correlation and mutual information especially for parity problems.

With the help of classification experiments on real world data we showed that the discrimination of positive interactions, synergies, and negative interactions, redundancy, is valuable in information fusion, which is necessary in multimedia classification and retrieval. Here especially synergistic feature sets prove to be beneficial to reduce the input feature set drastically, but this is for now going along with a loss in performance. This drawback we hope to overcome with the development

of an efficient and hence complete calculation of all relevant feature information interactions. Other directions in future work is to apply more sophisticated feature selection and construction approaches.

Other directions of research will be the utilization of more complex multimedia data such as the Wikipedia collection and tests with more sophisticated features like moment-based visual features.

**Acknowledgement** This work is funded by the European project MultiMATCH (EU-IST-STREP#033104).

## References

1. Bell A (2003) The co-information lattice. In: 4th Int. symposium on independent component analysis and blind signal seperation (ICA2003), Nara, 1–4 April 2003, pp 921–926
2. Benitez AB, Chang SF (2002) Multimedia knowledge integration, summarization and evaluation. In: Workshop on multimedia data mining, Edmonton, 23–26 July 2002, pp 23–26
3. Bruno E, Moenne-Loccoz N, Marchand-Maillet S (2008) Design of multimodal dissimilarity spaces for retrieval of multimedia documents. IEEE Trans Pattern Anal Mach Intell 30(9): 1520–1533
4. Chechik G, Tishby N (2002) Extracting relevant structures with side information. Adv Neural Inf Process Syst 15:857–864
5. Dass S, Jain A, Nandakumar K (2005) A principled approach to score level fusion in multimodal biometric systems. In: Proceedings of audio- and video-based biometric person authentication (AVBPA), Hilton Rye Town, 20–22 July 2005, pp 1049–1058
6. Fleuret F (2004) Fast binary feature selection with conditional mutual information. Mach Learn Res 5:1531–1555
7. Freitas AA (2001) Understanding the crucial role of attribute interaction in data mining. Artif Intell Rev 16(3):177–199
8. Hyvarinen A, Oja E (2000) Independent component analysis: algorithms and applications. Neural Netw 13(4–5):411–430
9. Jajuga K, Papla D (2006) Copula functions in model based clustering. Stud Classif Data Anal Knowl Organ 15:39–55
10. Jakulin A, Bratko I (2003) Analyzing attribute dependencies. In: Principles of knowledge discovery in data (PKDD), Cavtat-Dubrovnik, 22–26 September 2003, pp 229–240
11. Joachims T (2002) Learning to classify text using support vector machines. Kluwer, Boston
12. Kokar M, Weyman J, Tomasik J (2004) Formalizing classes of information fusion systems. Inf Fusion 5:189–202
13. Kolenda T, Winther O, Hansen L, Larsen J (2002) Independent component analysis for understanding multimedia content. Neural Netw Signal Process 12:757–766
14. Kononenko I, Simec E, Robnik-Sikonja M (1997) Overcoming the myopia of inductive learning algorithms with relieff. Appl Intell 7(1):39–55
15. Liu T, Chen Z, Zhang B, Ma W, Wu G (2004) Improving text classification using local latent semantic indexing. In: Fourth IEEE international conference on data mining (ICDM'04), Brighton, 1–4 November 2004, pp 162–169
16. Nemenman I (2004) Information theory, multivariate dependence and genetic networks. Technical Report NSF-KITP-04-54, KITP, UCSB
17. Pazzani M (1996) Searching for dependencies in bayes classifiers. In Fisher D, Lenz H-J (eds) Learning from data: AI and statistics. Springer, Heidelberg
18. Perez I (1997) Learning in presence of complex attribute interactions: an approach based on relational operators. PhD thesis
19. Poh N, Bengio S (2005) How do correlation and variance of base-experts affect fusion in biometric authentication tasks? In: IEEE transactions on acoustics, speech, and signal processing, vol 53. IEEE, Piscataway, pp 4384–4396
20. Tumer K, Gosh J (1999) Linear order statistics combiners for pattern classification. In: Combining artificial neural networks

21. Vasconcelos N, Carneiro G (2002) What is the role of independence for visual recognition? In: European conference on computer vision, Copenhagen, 27 May–2 June 2002, pp 297–311
22. Vinokurow A, Hardoon D, Shawe-Taylor J (2003) Learning the semantics of multimedia content with application to web image retrieval and classification. In: Fourth international symposium on independent component analysis and blind source separation, Nara, April 2003
23. Westerveld T, de Vries AP (2004) Multimedia retrieval using multiple examples. In: International conference on image and video retrieval (CIVR'04), Dublin, July 2004
24. Westerveld T, van Zwol R (2007) Multimedia retrieval at inex 2006. In: ACM SIGIR forum, vol 41(1). ACM, New York, pp 58–63
25. Wu L, Cohen P, SL O (2002) From members to team to committee—a robust approach to gestural and multimodal recognition. Trans Neural Netw 13:972–982
26. Wu Y, Chang EY, Smith JR (2004) Optimal multimodal fusion for multimedia data analysis. In: MULTIMEDIA '04: proceedings of the 12th annual ACM international conference on multimedia. ACM, New York, pp 572–579. doi:http://doi.acm.org/10.1145/1027527.1027665
27. Yan R, Hauptmann A (2003) The combination limit in multimedia retrieval. In: MULTIMEDIA '03: proceedings of the eleventh ACM international conference on multimedia. ACM, New York, pp 339–342
28. Zhao R, Grosky W (2002) Narrowing the semantic gap—improved text-based web document retrieval using visual features. In: IEEE transactions on multimedia, vol 4(2). IEEE, Piscataway, pp 189–200

**Jana Kludas** received her Diplom-Ingenieur (equivalent to master degree) at Technical University Ilmenau, Germany in 2006. She is currently a Research Assistant with the Computer Vision and Multimedia Laboratory, University of Geneva, Switzerland. Her research interests are image processing, pattern recognition and information fusion. She is working currently on her Ph.D. thesis focused on retrieval and classification of multimedia documents.

**Eric Bruno** received his M.S. degree from the Engineers School of Physics in Strasbourg, France in 1995, and his Ph.D in signal processing from the Joseph Fourier University, Grenoble, France in 2001. Since 2002, he is working at the Computer Vision and Multimedia Laboratory, University of Geneva, Switzerland, as a research associate. His research interests focus on video analysis, content-based multimedia retrieval and multimodal information fusion.

**Stéphane Marchand-Maillet** received his PhD on theoretical image processing from Imperial College, London in 1997. He then joined the Institut Eurecom at Sophia-Antipolis (France) where he worked on automatic video indexing techniques based on human face localization and recognition. Since 1999, he is Assistant Professor in the Computer Vision and Multimedia Lab at the University of Geneva, where he is working on content-based multimedia retrieval as head of the *Viper* research group. The group is involved in several national and international research projects. He has authored several publications on multimedia information retrieval and related fields, including a book on low-level image analysis.